Phylogenetic Networks from Partial Trees

S. Grünewald
Department of Combinatorics and Geometry (DCG),
CAS-MPG Partner Institute for Computational Biology (PICB),
Shanghai Institutes for Biological Sciences (SIBS),
Chinese Academy of Sciences (CAS), China,
and
K. T. Huber and **Q. Wu**
School of Computing Sciences,
University of East Anglia,
Norwich, NR5 7TJ,
United Kingdom.

February 1, 2008

K. T. Huber
School of Computing Sciences,
University of East Anglia,
Norwich, NR5 7TJ,
United Kingdom.
email: katharina.huber@cmp.uea.ac.uk
FAX: +44 (0) 1603 593345

We would prefer to submit the final manuscript in LATEX.

Abstract

A contemporary and fundamental problem faced by many evolutionary biologists is how to puzzle together a collection $\mathcal{P}$ of partial trees (leaf-labelled trees whose leaves are bijectively labelled by species or, more generally, taxa, each supported by e. g. a gene) into an overall parental structure that displays all trees in $\mathcal{P}$. This already difficult problem is complicated by the fact that the trees in $\mathcal{P}$ regularly support conflicting phylogenetic relationships and are not on the same but only overlapping taxa sets. A desirable requirement on the sought after parental structure therefore is that it can accommodate the observed conflicts. Phylogenetic networks are a popular tool capable of doing precisely this. However, not much is known about how to construct such networks from partial trees, a notable exception being the Z-closure super-network approach and the recently introduced Q-imputation approach. Here, we propose the usage of closure rules to obtain such a network. In particular, we introduce the novel Y-closure rule and show that this rule on its own or in combination with one of Meacham's closure rules (which we call the M-rule) has some very desirable theoretical properties. In addition, we use the M- and Y-rule to explore the dependency of Rivera et al.'s "ring of life" on the fact that the underpinning phylogenetic trees are all on the same data set. Our analysis culminates in the presentation of a collection of induced subtrees from which this ring can be reconstructed.

1 Introduction

Phylogenetic trees have proved an important tool for representing evolutionary relationships. For a set X of species (or, more generally, taxa) these are formally defined as leaf-labelled trees whose leaves are bijectively labelled by the elements of X. Advances in DNA sequencing have resulted in ever more data on which such trees may be based. Computational limitations however combined with the need to understand species evolution have left biologists with the following fundamental problem which we will refer to as *amalgamation problem*: given a collection $\mathcal{P}$ of phylogenetic trees, how can these trees be amalgamated into an overall parental structure that preserves the phylogenetic relationships supported by the trees in $\mathcal{P}$? The hope is that such a structure might help shed light on the evolution of the underlying genomes (and thus the species).

In the ideal case that all trees in $\mathcal{P}$ support the same phylogenetic relationships (as is the case for trees $\mathcal{T}_1$ and $\mathcal{T}_2$ depicted in Fig. 1) this structure is known to be a phylogenetic tree and a supertree method [2] may be used to reconstruct it. For the above example the outcome $\mathcal{T}^*$ of such a method is
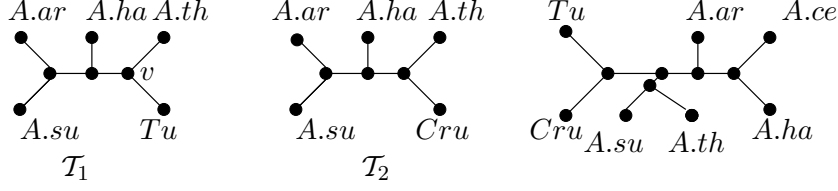
Figure 1: 3 phylogenetic trees which appeared in weighted form in [17] on subsets of the 7 plant species: *A.thaliana (A.th)*, *A.suecia (A.su)*, *Turritis (Tu) A.arenosa (A.ar)*, *A.cebennensis (A.ce)*, *Crucihimalaya (Cru)* and *A.halleri (A.ha)*.

$\mathcal{T}_1$ with species *Cru* (see Fig. 1 for full species names) attached via a pendant edge to the vertex labelled v. It should be noted that $\mathcal{T}^*$ supports the same phylogenetic relationships as $\mathcal{T}_1$ and $\mathcal{T}_2$ in the following sense: For a finite set X, call a bipartition $S = \{A, \widetilde{A}\}$ of some subset $X' \subseteq X$ a *partial split* on X, or a *partial (X)-split* for short, and denote it by $A|\widetilde{A}$ or, equivalently, by $\widetilde{A}|A$ where $\widetilde{A} := X' - A$. In particular, call S a *(full) split* of X if $X' = X$. Furthermore, say that a partial X-split $S = A|\widetilde{A}$ *extends* a partial X-split $S' = B|\widetilde{B}$ if either $B \subseteq A$ and $\widetilde{B} \subseteq \widetilde{A}$ or $B \subseteq \widetilde{A}$ and $A \subseteq \widetilde{B}$. Finally, say that a phylogenetic tree $\mathcal{T}$ *displays* a split $S = A|\widetilde{A}$ if S is a partial split on the leaf set $\mathcal{L}(\mathcal{T})$ of $\mathcal{T}$ induced by deleting an edge of $\mathcal{T}$. Then "supports the same phylogenetic relationships" means that for every split S displayed by $\mathcal{T}_1$ or $\mathcal{T}_2$ there exists a split on $\mathcal{L}(\mathcal{T}^*)$ that extends S and is displayed by $\mathcal{T}^*$.

Due to complex evolutionary mechanisms such as incomplete lineage sorting, recombination (in the case of viruses), or lateral gene transfer (in case of bacteria) the trees in $\mathcal{P}$ may however support not the same but conflicting phylogenetic relationships. A phylogenetic network in the form of a split network (see [10, 19] for overviews) rather than a phylogenetic tree is therefore the structure of choice if one wishes to simultaneously represent all phylogenetic relationships supported by the trees in $\mathcal{P}$. An example in point is the split network pictured in Fig. 2 which appeared as a weighted network in [17]. With replacing "edge" in the definition of displaying by "band of parallel edges" and "$\mathcal{L}(\mathcal{T})$" by "set of network vertices of degree 1" to obtain a definition for when a split network displays a split, it is straight forward to check that the network in Fig. 2 displays all splits displayed by the 3 trees pictured in Fig. 1.

It should be noted that phylogenetic networks such as the one depicted in

Fig. 2 (see e.g. [7, 11, 14] for recently introduced other types of phylogenetic networks) provide a means to visualize the complexity of a data set and should not be thought of as an explicit model of evolution. Awareness of this complexity does not only allow the exploration of a data set but, as is the case of e.g. hybridization networks [14], can also serve as starting point for obtaining an explicit model of evolution (see [12] for more on this).
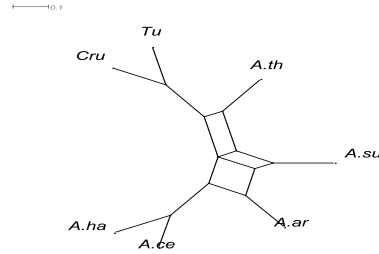

Figure 2: A circular phylogenetic network that represents all phylogenetic relationships supported by the trees depicted in Fig. 1 (see that figure for full species names).

Apart from displaying all splits induced by the 3 trees depicted in Fig. 1, the network depicted in Fig. 2 has a further interesting feature. It is *circular*. In other words, if X denotes the set of the 7 plant species under consideration, then the elements of X can be arranged around a circle C so that every split $S = A|\widetilde{A}$ of X displayed by the network can be obtained by intersecting C with a straight line so that the label set of one of the resulting 2 connected components is A and the label set of the other is $\widetilde{A}$.

Although seemingly a very special type of phylogenetic network, circular phylogenetic networks are a frequently used structure in phylogenetics (see e.g. [3, 4, 5, 6, 8]) as they do not only naturally generalize the concept of a phylogenetic tree but are also guaranteed to be representable in the plane; a fact that greatly facilitates drawing and thus analyzing them. However, although recently first steps have been made with regards to finding a solution to the amalgamation problem in terms of a phylogenetic network leading to the attractive Z-closure [13] and Q-imputation [9] approaches, very little is known about a solution of this problem in terms of a circular phylogenetic network.

Intrigued by this and motivated by the fact that, from a combinatorial point of view, phylogenetic trees and networks are *split systems* (i. e. collections

of full splits) and that therefore the amalgamation problem boils down to the problem of how to extend partial splits on some set X to splits on X, we wondered whether closure rules for partial splits could not be of help. Essentially mechanisms for splits' enlargement, such rules have proved useful for supertree construction and also underpin the above mentioned Z-closure super-network approach. As it turns out, this is indeed the case. As an immediate consequence of our main result (Corollary 5.5), we obtain that for a collection of partial splits that can be "displayed" by a circular phylogenetic network $\mathcal{N}$, the collection of (full) splits generated by the closure rules in the centre of this paper is guaranteed to be displayable by $\mathcal{N}$ and also independent of the order in which the rules are applied.

In a study aimed at shedding light into the origin of eukaryotes, Rivera et al. [20] put forward the idea of a "ring of life" with the eukaryotic genome being the result of a fusion of two diverse procaryotic genomes (see also [16, 20, 23]). A natural and interesting question in this context is how dependent Rivera et al.'s ring of life is on the fact that all underpinning trees are on the same taxa set. In the last section of this paper, we provide a partial answer by presenting an example of a collection of induced partial trees from which the ring of life can be reconstructed using the M- and Y-rule.

The paper is organized as follows. In Section 2, we first introduce some more terminology and then restate one of Meacham's closure rules (our M-rule) and introduce the novel Y-rule. In Section 3, we study the relationship between the M- and Y-rule and the closure rule that underpins the aforementioned Z-closure super-network approach. In Section 4, we introduce the concept of a circular collection of partial splits and show that both the Y- and M-rule preserve circularity (Proposition 4.4). In Section 5, we introduce the concept of a split closure and show that for certain collections of partial splits this closure is independent of the order in which the Y-rule and/or M-rule are/is applied (Theorem 5.3). This result lies at the heart of Corollary 5.5. In Section 6, we explore the dependency of Rivera et al.'s ring of life on the fact that the underpinning trees are all on the same data set

Throughout the paper, X denotes a finite set and the terminology and notation largely follows [21].

6

2 Closure rules

We start this section by introducing some additional terminology and notation. Subsequent to this, we first restate Meacham's rule (which we call the M-rule) and then introduce a novel closure rule which we call the Y-rule.

Let $\Sigma(X)$ denote the collection of all partial splits of X and suppose $\Sigma \subseteq \Sigma(X)$. Then a partial split $S \in \Sigma$ that can be extended by a partial split $S' \in \Sigma - S$ is called *redundant*. The set obtained by removing redundant elements from Σ is denoted by Σ^-. If $\Sigma = \Sigma^-$ then Σ is called *irreducible* and the set of all irreducible subsets in $\Sigma(X)$ is denoted by $\mathcal{P}(X)$. Note that the relation "$\preceq$" defined for any two (partial) split collections $\Sigma, \Sigma' \in \mathcal{P}(X)$ by putting $\Sigma \preceq \Sigma'$ if every partial split in Σ is extended by a partial split in Σ' is a partial order on $\mathcal{P}(X)$ [21].

Suppose for the following that θ is a closure rule, that is, a replacement rule that replaces a collection $\mathcal{A} \subseteq \Sigma(X)$ of partial splits that satisfy some condition C_θ by a collection $\theta(\mathcal{A}) \subseteq \Sigma(X)$ whose elements are generated in some systematic way from the partial splits in $\mathcal{A}$ (see e.g. the M- and the Y-closure rules presented below for two such systematic ways). Suppose $\Sigma, \Sigma' \in \mathcal{P}(X)$ are two irreducible collections of partial splits and $C_\theta(\Sigma)$ is the set of all subsets of Σ that satisfy C_θ. If there exists some subset $\mathcal{A} \in C_\theta(\Sigma)$ such that $\Sigma' = (\Sigma \cup \theta(\mathcal{A}))^-$ then we say that Σ' is *obtained from Σ via a single application of θ*. Finally, if for every subset $\mathcal{A} \in C_\theta(\Sigma)$ we have $\theta(\mathcal{A})^- \preceq \Sigma$ then we call an application of θ to Σ *trivial* and say that Σ is *closed* with respect to θ.

We are now in the position to present the 2 closure rules we are mostly concerned with in this paper: the M-rule which is originally due to Meacham [18] and the novel Y-rule. We start with Meacham's rule.

2.1 The M-rule

Suppose $S_1, S_2 \in \Sigma(X)$ are two distinct partial splits of X. Then the *M-rule* θ_M is as follows:

(θ_M) If there exists some $A_i \in S_i$, $i = 1, 2$ such that

$$A_1 \cap A_2 \neq \emptyset \text{ and } \widetilde{A_1} \cap \widetilde{A_2} \neq \emptyset \tag{1}$$

then replace $\mathcal{A} = \{S_1, S_2\}$ by the set $\theta_M^{\{A_1, A_2\}}(\mathcal{A})$ which comprises of $\mathcal{A}$ and, in addition, also the partial splits

$$S_1' = (A_1 \cap A_2)|(\widetilde{A_1} \cup \widetilde{A_2}) \text{ and } S_2' = (\widetilde{A_1} \cap \widetilde{A_2})|(A_1 \cup A_2).$$

In case the partial splits S_1 and S_2 are such that there is no ambiguity with regards to the identity of the sets A_1 and A_2 in the statement of the M-rule or they are irrelevant to the discussion, we will simplify $\theta_M^{\{A_1,A_2\}}(\mathcal{A})$ to $\theta_M(\mathcal{A})$. Clearly, such ambiguity cannot arise if S_1 and S_2 are *compatible*, that is, there exist subsets $D_i \in S_i$, $i = 1, 2$ such that $D_1 \cap D_2 = \emptyset$. However if S_1 and S_2 are *incompatible*, that is, not compatible then caution is required.

Note that if A_1 and A_2 as in the statement of the M-rule are such that $A_2 \subseteq A_1$ and $\widetilde{A_1} \subseteq \widetilde{A_2}$, then it is easy to verify that θ_M applies trivially to $\mathcal{A}$. Also note that for any $\Sigma \in \mathcal{P}(X)$ and any two distinct partial splits $S_1, S_2 \in \Sigma$, we have

$$\Sigma \preceq (\Sigma \cup \theta_M(\{S_1, S_2\}))^-.$$

Finally, note that any phylogenetic tree on X that displays the partial splits in some set $\Sigma \in \mathcal{P}(X)$ also displays the partial splits in $(\Sigma \cup \theta_M(\{S_1, S_2\}))^-$, $S_1, S_2 \in \Sigma$.

2.2 The Y-rule

Suppose $S_i \in \Sigma(X)$, $i = 1, 2, 3$, are three distinct partial splits of X. Then the Y-*rule* θ_Y is as follows:

(θ_Y) If there exists some $A_i \in S_i$, $i = 1, 2, 3$ such that

$$\emptyset \notin \{A_1 \cap A_2 \cap A_3, \widetilde{A_1} \cap \widetilde{A_2} \cap A_3, \widetilde{A_1} \cap A_2 \cap \widetilde{A_3}\} \text{ and}$$
$$A_1 \cap \widetilde{A_2} \cap \widetilde{A_3} = \emptyset. \tag{2}$$

(see Fig. 3(a) for a graphical interpretation), then replace $\mathcal{A} = \{S_1, S_2, S_3\}$ by the set $\theta_Y^{\{A_1,A_2,A_3\}}(\mathcal{A})$ which comprises of the partial splits

$$
\begin{aligned}
S_1' &= \widetilde{A_1} \cup (\widetilde{A_2} \cap \widetilde{A_3})|A_1, S_2' = A_2 \cup (A_1 \cap \widetilde{A_3})|\widetilde{A_2}, \text{ and} \\
S_3' &= A_3 \cup (A_1 \cap \widetilde{A_2})|\widetilde{A_3}.
\end{aligned}
$$

Although the condition in (2) might look quite strange at first sight, the class of triplets of partial splits that satisfy it is very rich. For example, suppose that $S_i = A_i|\widetilde{A_i}$, $i = 1, 2, 3$ are splits of X that can be arranged in the plane as indicated in Fig. 3(b) where each bold, straight line represents one of S_i, $i = 1, 2, 3$ and the dots represent non-empty triplewise intersections of the parts of S_i, $i = 1, 2, 3$, in which they lie. For example, the dot in the bottom wedge represents the intersection $A_1 \cap A_2 \cap A_3$. The shaded regions correspond to the 3 non-empty intersections mentioned in the statement of the Y-rule. The partial splits $S_i' = A_i'|\widetilde{A_i'}$ $i = 1, 2, 3$ obtained by restricting

$$\widetilde{A_1} \cap \widetilde{A_2} \cap A_3 \neq \emptyset \qquad \widetilde{A_1} \cap A_2 \cap \widetilde{A_3} \neq \emptyset$$

$$A_1 \cap \widetilde{A_2} \cap \widetilde{A_3} = \emptyset$$

$$A_1 \cap A_2 \cap A_3 \neq \emptyset$$

(a)

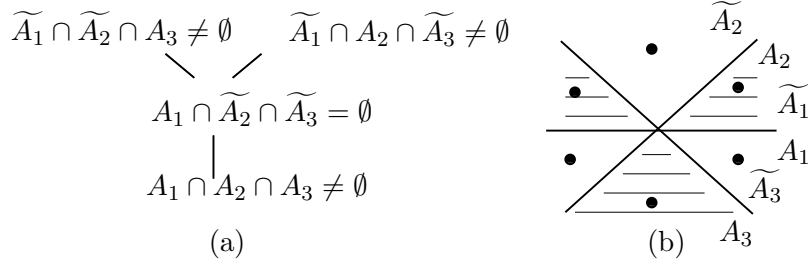$\widetilde{A_2}$, A_2, $\widetilde{A_1}$, A_1, $\widetilde{A_3}$, A_3

(b)

Figure 3: (a) A graphical representation of Condition (2) in the form of a Y. (b) An example of three splits, depicted in bold lines, that satisfy Condition (2) – see text for details.

S_1, S_2, and S_3 to different subsets of X so that the shaded regions remain non-empty form a triplet of partial splits that satisfy (2).

As the example of set $\mathcal{A}$ comprising the three partial splits $S_1 = 145|2367$, $S_2 = 1357|246$, and $S_3 = 127|356$ shows different choices of the sets A_i, $i = 1, 2, 3$ lead to different sets $\theta_Y^{\{A_1, A_2, A_3\}}(\mathcal{A})$. For example, if $A_1 := \{1, 4, 5\}$, $A_2 := \{1, 3, 5, 7\}$, and $A_3 := \{1, 2, 7\}$ then (2) is satisfied and $\theta_Y^{\{A_1, A_2, A_3\}}(\mathcal{A}) = \{S_1, S_2, 1247|356\}$. If however A_1 and A_2 are as before and $A_3 := \{3, 5, 6\}$, then (2) is also satisfied and $\theta_Y^{\{A_1, A_2, A_3\}}(\mathcal{A})$ is the set $\{S_1, S_2, 127|3456\}$. Following our practise for the M-rule, for $\mathcal{A} = \{S_1, S_2, S_3\}$ we simplify $\theta_Y^{\{A_1, A_2, A_3\}}(\mathcal{A})$ to $\theta_Y(\mathcal{A})$ if the partial splits S_i, $i = 1, 2, 3$ are such that there is no ambiguity with regards to the identity of the sets A_i, $i = 1, 2, 3$, in the statement of the Y-rule or they are irrelevant to the discussion.

Note that if A_i, $i = 1, 2, 3$ as in the statement of the Y-rule are such that, in addition, $\emptyset \neq A_1 \cap \widetilde{A_2} \subseteq A_3$, $\emptyset \neq A_1 \cap \widetilde{A_3} \subseteq A_2$, $\emptyset \neq \widetilde{A_3} \cap \widetilde{A_2} \subseteq \widetilde{A_1}$, and $\widetilde{A_1} \cap A_2 \cap A_3 \neq \emptyset$ it is easy to see that θ_Y applies trivially to $\mathcal{A}$. Also note that for any $\Sigma \in \mathcal{P}(X)$ and any 3 partial splits $S_1, S_2, S_3 \in \Sigma$ of X, we have

$$\Sigma \preceq (\Sigma \cup \theta_Y(\{S_1, S_2, S_3\}))^-.$$

3 First closure rule relationships

In this section we first restate the Z-(closure) rule which was used in [13] in the context of a supernetwork construction approach and then investigate the relationship between the Y-, M-, and Z-rule.

Also originally due to Meacham [18], the Z-rule θ_Z can be restated as

follows: Suppose $S_1, S_2 \in \Sigma(X)$ are two distinct partial splits of X.

(θ_Z) If there exists some $A_i \in S_i$, $i = 1, 2$ such that

$$\emptyset \notin \{A_1 \cap A_2, A_2 \cap \widetilde{A_1}, \widetilde{A_1} \cap \widetilde{A_2}\} \text{ and } A_1 \cap \widetilde{A_2} = \emptyset \qquad (3)$$

then replace $\mathcal{A} = \{S_1, S_2\}$ by the set $\theta_Y(\mathcal{A})$ which comprises of the partial splits $(\widetilde{A_1} \cup \widetilde{A_2})|A_1$ and $\widetilde{A_2}|(A_1 \cup A_2)$.

Note that any two compatible partial splits of X satisfy the condition in (3).

With this third closure rule at hand we are now in the position to present a first easy to verify result. Suppose S_1, S_2, and S_3 are 3 distinct partial splits of X such that there exist parts $A_i \in S_i$, $i = 1, 2, 3$ as in the statement of the Y-rule. If, in addition, $\widetilde{A_1} \cap \widetilde{A_2} \cap \widetilde{A_3} \neq \emptyset$ and $A_1 \subseteq A_2 \cup A_3$, then the partial split $A_1|\widetilde{A_1} \cup (\widetilde{A_2} \cap \widetilde{A_3})$ generated by θ_Y is also generated by first applying θ_M to S_2 and S_3 (with regards to $\widetilde{A_2} \cap \widetilde{A_3} \neq \emptyset$) and then applying θ_M to the resulting partial split $A_2 \cup A_3|\widetilde{A_2} \cap \widetilde{A_3}$ and S_1.

In addition, we have the following result whose straight forward proof we leave to the reader.

Proposition 3.1 *Suppose S_1 is a full split of X. Then the following statements hold.*

(i) *If S_2 is a partial X-split and θ_Z applies to $\Sigma = \{S_1, S_2\}$, then*

$$\theta_M(\Sigma)^- = \theta_Z(\Sigma).$$

(ii) *If S_2 and S_3 are partial X-splits so that θ_Y applies to $\Sigma = \{S_1, S_2, S_3\}$ and θ_Z applies to $\{S_1, S_2\}$ and $\{S_1, S_3\}$. Then*

$$(\theta_Y(\Sigma) \cup \bigcup_{j=2,3} \theta_M(S_1, S_j))^- = (\bigcup_{i \in \{2,3\}} \theta_Z(S_1, S_i))^-.$$

4 Closure rules and weakly compatible collections of partial splits

In this section we introduce the notion of a weakly compatible collection of partial splits and study properties of the Y- and M-rules regarding such collections. A particular focus lies on the study of circular collections of partial splits which we also introduce. As we will see, they form a very rich subclass of such collections of partial splits.

4.1 Weakly compatible collections of partial splits

We start this section with a definition that generalizes the concept of weak compatibility for (full) splits of X [1] to partial splits of X. Suppose $S_i = A_i|\widetilde{A_i} \in \Sigma(X)$, $i = 1, 2, 3$, are three partial X-splits. Then we call S_1, S_2, S_3 *weakly compatible* if at least one of the four intersections

$$A_1 \cap A_2 \cap A_3, \widetilde{A_1} \cap \widetilde{A_2} \cap A_3, \widetilde{A_1} \cap A_2 \cap \widetilde{A_3}, A_1 \cap \widetilde{A_2} \cap \widetilde{A_3} \qquad (4)$$

is empty[1]. Since the roles of A_i and $\widetilde{A_i}$ in S_i, $i = 1, 2, 3$, can be interchanged without changing S_1, S_2, S_3 we have that S_1, S_2, S_3 are weakly compatible if and only if at least one of the four intersections

$$\widetilde{A_1} \cap \widetilde{A_2} \cap \widetilde{A_3}, A_1 \cap A_2 \cap \widetilde{A_3}, A_1 \cap \widetilde{A_2} \cap A_3, \widetilde{A_1} \cap A_2 \cap A_3$$

is empty. More generally, we call a collection $\Sigma \subseteq \Sigma(X)$ of partial X-splits *weakly compatible* if every three partial splits in Σ are weakly compatible. To give an example, the partial splits $S_1 = 123|4567$, $S_2 = 124|3567$, and $S_3 = 235|146$ are weakly compatible whereas the partial splits S_3, $S_4 = 24|135$, and $S_5 = 21|346$ are not. Thus, $\{S_1, \ldots, S_5\}$ is not weakly compatible. Note that, like in the case of (full) splits, it is easy to see that any collection of pairwise compatible partial splits is also weakly compatible.

Clearly any three partial splits $S_i = A_i|\widetilde{A_i} \in \Sigma(X)$, $i = 1, 2, 3$, for which precisely one of the four intersections in (4) is empty also satisfies Condition (2). Thus θ_Y may be applied to S_1, S_2, S_3. However, as the example of the set $\{127|3456, 1234|567, 235|146\}$ shows, application of θ_Y to a weakly compatible collection of partial splits does not, in general, yield a weakly compatible collection of partial splits. Also it should be noted that θ_M applied to a weakly compatible collection of partial splits does not always yield a weakly compatible collection of partial splits.

However, the next result whose proof is straight forward holds.

Lemma 4.1 *Suppose* $\Sigma, \Sigma' \subseteq \Sigma(X)$. *If* Σ' *is weakly compatible and* $\Sigma \preceq \Sigma'$, *then* Σ *must also be weakly compatible.*

4.2 Circular collections of partial splits

We now turn our attention to the study of a special class of weakly compatible collections of partial splits called circular collections of partial splits. To

---

[1]In the definition of weak compatibility for full splits, S_1, S_2 and S_3 are full splits and the condition in (4) is the same (see [1]).

be able to state their definition, we require some more terminology which we introduce next.

A *cycle* C is a connected graph with $|V(C)| \geq 3$ and every vertex has degree 2. We call C an *X-cycle* if the vertex set of C is X. For $x_i \in X$ $(1 \leq i \leq n := |X|)$ and C an X-cycle, we call $x_1, x_2, \ldots, x_n, x_{n+1} = x_1$ a *vertex ordering (of C)* if the edge set of C coincides with the set of all 2-sets $\{x_i, x_{i+1}\}$ of X, $i = 1, \ldots, n$.

For a graph $G = (V, E)$ and some subset E' of E, we denote by $G - E'$ the graph obtained from G by deleting the edges in E'. We say that a partial X-split $A | \widetilde{A}$ is *displayed* by an X-cycle C if there exist two distinct edges e_1 and e_2 in C such that the vertex set of one of the two components of $C - \{e_1, e_2\}$ contains A and the other one contains $\widetilde{A}$. More generally, we say that a set $\Sigma \subseteq \Sigma(X)$ of partial splits is *displayed* by an X-cycle C if every partial split in Σ is displayed by C. Finally, we say that a collection $\Sigma \subseteq \Sigma(X)$ is *circular* if there exists some X-cycle C such that every partial split in Σ is displayed by C. Note that every split collection in $\Sigma(X)$ displayed by a circular phylogenetic network is circular.

As is well-known, every circular split system is in particular weakly compatible. The next result shows that an analogous result holds for collections of partial splits.

Lemma 4.2 *Suppose $\Sigma \subseteq \Sigma(X)$. If Σ is circular then Σ is also weakly compatible.*

Proof: Suppose C is an X-cycle that displays Σ but there exist three partial splits $S_1, S_2, S_3 \in \Sigma$ such that with $A_i \in S_i$, $i = 1, 2, 3$, playing the role of their namesakes in (4) none of the four intersections in (4) is empty. Then S_1 and S_2 are incompatible and, since S_1 and S_2 are displayed by C, there must exist edges $e_1, e_1', e_2, e_2' \in E(C)$ such that, for all $i, j \in \{1, 2\}$ distinct, the vertex set of one component of $C - \{e_i, e_i'\}$ contains $A_i \cup e_j$ and the other contains $\widetilde{A_i} \cup e_j'$. Since S_3 is displayed by C and neither $A_1 \cap A_2 \cap A_3$ nor $\widetilde{A_1} \cap \widetilde{A_2} \cap A_3$, nor $\widetilde{A_1} \cap A_2 \cap \widetilde{A_3}$ is empty, it follows that $A_1 \cap \widetilde{A_2} \cap \widetilde{A_3} = \emptyset$, which is impossible. ∎

As in the case of full splits, the converse of the above lemma is not true in general. For example, the set Σ comprising the partial splits $S_1 = 12|35$, $S_2 = 125|34$, $S_3 = 13|245$ and $S_4 = 135|24$ is weakly compatible since the sets $\{S_1, S_2\}$ and $\{S_3, S_4\}$ are pairwise compatible. Yet, as can be easily checked, Σ is not circular.

Corresponding to Lemma 4.1, we have:

Lemma 4.3 *Suppose $\Sigma, \Sigma' \subseteq \Sigma(X)$. If Σ' is displayed by an X-cycle C and $\Sigma \preceq \Sigma'$, then Σ is also displayed by C.*

4.3 Circularity and the M- and Y-rule

As was noted earlier, neither the Y-rule nor the M-rule preserve weak compatibility in general. As the next result shows, the situation is different for the special case of circular collections of partial splits.

Proposition 4.4 *Suppose $\Sigma, \Sigma' \in \mathcal{P}(X)$ and C is an X-cycle. If Σ' is obtained from Σ by a single application of either θ_Y or θ_M then Σ is displayed by C if and only if Σ' is displayed by C.*

Proof: Suppose $\Sigma, \Sigma' \in \mathcal{P}(X)$ and C is an X-cycle. We start the proof with noting that, regardless of whether Σ' is obtained from a single application of either θ_Y or θ_M to Σ, Σ is displayed by C whenever Σ' is displayed by C in view of Lemma 4.3.

Conversely, suppose that Σ is displayed by C. Assume first that Σ' is obtained from Σ by a single application of θ_Y. Let $\{S_1, S_2, S_3\} \subseteq \Sigma$ be the set to which θ_Y is applied. With $A_i \in S_i$, $i = 1, 2, 3$, playing the role of their namesakes in the statement of (2), we may assume without loss of generality that none of the three intersections $D_1 = A_1 \cap A_2 \cap A_3$, $D_2 = \widetilde{A_1} \cap \widetilde{A_2} \cap A_3$, and $D_3 = \widetilde{A_1} \cap A_2 \cap \widetilde{A_3}$ is empty but that $A_1 \cap A_2 \cap \widetilde{A_3} = \emptyset$. It suffices to show that the partial split $S = A_3 \cup (A_1 \cap A_2) | \widetilde{A_3}$ is displayed by C.

Clearly, if $A_1 \cap \widetilde{A_2} = \emptyset$ then $S = S_3$ and, therefore, S is displayed by C. So assume $A_1 \cap \widetilde{A_2} \neq \emptyset$. Then since by assumption $D_i \neq \emptyset$, $i = 1, 2, 3$, and S_1 and S_2 are displayed by C, there must exist four distinct edges $e_1, e_1', e_2, e_2' \in E(C)$ such that, for all $i, j \in \{1, 2\}$ distinct, one component of $C - \{e_i, e_i'\}$ contains A_i in its vertex set and $e_j \subseteq A_i$ and the other contains $\widetilde{A_i}$ in its vertex set and $e_j' \subseteq \widetilde{A_i}$. Without loss of generality, we may assume that $X = \{x_1, \ldots, x_n\}$, $n \geq 3$, that $x_1, x_2, \ldots, x_n$ is a vertex ordering of C, and that $e_1 = \{x_n, x_1\}$. Furthermore, we may also assume without loss of generality that the component of $C - \{e_1, e_1'\}$ that contains x_1 in its vertex set also contains A_1. Since $D_1 \neq \emptyset \neq D_2$, and S_3 is displayed by C there must exist distinct paths P and P' in C such that either $\widetilde{A_3} \subseteq V(P)$ or $\widetilde{A_3} \subseteq V(P')$ (see Figure 4). If $\widetilde{A_3} \subseteq V(P)$ then

$$\emptyset = \widetilde{A_1} \cap A_2 \cap \widetilde{A_3} = D_3$$

which is impossible. Thus $\widetilde{A_3} \subseteq V(P')$ must hold. Suppose $y, z \in V(C)$ are such that when starting at x_1 and traversing C clockwise y is contained in

13

A_3 and the next vertex y' on C with $y' \in A_3 \cup \widetilde{A_3}$ is contained in $\widetilde{A_3}$ whereas $z \in \widetilde{A_3}$ and the next vertex z' on C with $z' \in A_3 \cup \widetilde{A_3}$ is contained in A_3.
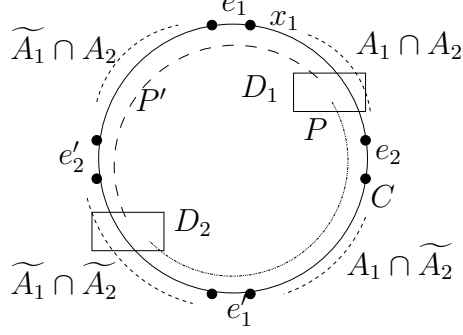

Figure 4: A schematic representation of the two alternative locations for $\widetilde{A_3}$ (cf proof of Proposition 4.4). The closed curve is the X-cycle C, the four curves with the short dashes represent the four non-empty intersections $A_1 \cap A_2$, $A_1 \cap \widetilde{A_2}$, $\widetilde{A_1} \cap \widetilde{A_2}$ and $\widetilde{A_1} \cap A_2$ (note that each of them can consist of more than one part), the rectangles mark the intersections D_1 and D_2, and the dotted and dashed curves represent the two paths P and P' on C on which $\widetilde{A_3}$ can lie.

Let P'' denote the path from z' to y (taken clockwise). Then e_2 and e_1' are edges on P'' and so $A_1 \cap \widetilde{A_2} \subseteq V(P'')$. The choice of y and z' implies $V(P'') \cap \widetilde{A_3} = \emptyset$ and $A_3 \cup (A_1 \cap \widetilde{A_2}) \subseteq V(P'')$. Hence, the split $V(P'')|X - V(P'')$ which is displayed by C extends the partial split S. Thus C displays S. This concludes the proof in case the applied closure rule applied is θ_Y.

To conclude the proof of the proposition suppose Σ' is obtained from Σ by a single application of θ_M. Let $\{S_1, S_2\} \subseteq \Sigma$ be the set to which θ_M is applied. With $A_i \in S_i$, $i = 1, 2$, we may assume without loss of generality that $A_1 \cap A_2 \neq \emptyset$ and $\widetilde{A_1} \cap \widetilde{A_2} \neq \emptyset$. If θ_M applies trivially to Σ then $\Sigma = \Sigma'$ and so Σ' must be displayed by C. If θ_M does not apply trivially to Σ it suffices to show that C displays $(A_1 \cap A_2)|(\widetilde{A_1} \cup \widetilde{A_2})$.

Since S_1 and S_2 are displayed by C there must exist edges $e_i, e_i' \in E(C)$ such that the vertex set of one of the two components P_i, P_i' of $C - \{e_i, e_i'\}$ contains A_i and the other contains $\widetilde{A_i}$, $i = 1, 2$. Put $k := |\{e_1, e_1'\} \cap \{e_2, e_2'\}|$ and note that $0 \leq k \leq 2$. Without loss of generality, we may assume $A_i \subseteq V(P_i)$ and $\widetilde{A_i} \subseteq V(P_i')$, $i = 1, 2$. Then, $\emptyset \neq A_1 \cap A_2 \subseteq V(P_1) \cap V(P_2)$. Since $V(P_1) \cap V(P_2)$ is the vertex set of one of the $4 - k$ components of C

14

with the edges e_i, e_i', $i = 1, 2$ removed, it follows that there must exist two distinct edges e_3, e_4 among the edges e_1, e_1', e_2, e_2' so that the vertex set of one of the two components of $C - \{e_3, e_4\}$ is $V(P_1) \cap V(P_2)$. Since

$$X - (V(P_1) \cap V(P_2)) = (X - V(P_1)) \cup (X - V(P_2)) = V(P_1') \cup V(P_2')$$

is the vertex set of the other component of $C - \{e_3, e_4\}$ and $\widetilde{A_i} \subseteq V(P_i')$, $i = 1, 2$, it follows that C displays $(A_1 \cap A_2)|(\widetilde{A_1} \cup \widetilde{A_2})$. This concludes the proof in case Σ' is obtained from Σ by a single application of θ_M and thus the proof of the proposition. ∎

Interestingly, the Z-rule does not preserve circularity in general. An example in point is the X-cycle C with $X = \{1, \dots, 5\}$ and the natural ordering of the elements of X as vertex ordering. Then the partial splits $S_1 = 13|45$ and $S_2 = 34|25$ are clearly displayed by C. Yet the Z-rule applied to $\{S_1, S_2\}$ generates the partial splits $13|245$ and $25|134$ which cannot be displayed by C.

5 Split closure sequences and split closures

In this section, we associate to a set Σ of partial splits a split closure sequence and define the last element of such a sequence to be a split closure of Σ. We also establish a key result for this paper which shows that under certain circumstances a split closure is unique.

5.1 Split closure sequences

Suppose $\Sigma \in \mathcal{P}(X)$ is a collection of partial splits that satisfies some partial splits property (P) such as, for example, weak compatibility and θ is one of the closure rules considered in this paper. Following [21], we associate a split closure sequence σ and a split closure to Σ as follows.

$$\sigma : \Sigma_0, \Sigma_1, \Sigma_2, \dots, \Sigma_i, \Sigma_{i+1}, \dots$$

is a strictly increasing (with respect to $\preceq$) sequence of sets in $\mathcal{P}(X)$ so that $\Sigma = \Sigma_0$ and, for all $i \geq 1$, Σ_{i+1} is obtained by one non-trivial application of θ to Σ_i whenever Σ_i satisfies (P). Note that since X is finite, there must exist a last element Σ_n in σ such that Σ_n either satisfies (P) and is closed under θ or Σ_n does not satisfy (P). In the latter case we reset Σ_n to be a new element $\omega \notin \mathcal{P}(X)$. We refer to σ as a *split closure sequence* for Σ

and call n the length of σ. In addition, we call the last element of σ a *split closure* of Σ. Note that in case $\Sigma_n \neq \omega$, θ applies only trivially to Σ_n.

The following combinations of (P) and θ are of interest to us:

(a) (P) is the property that Σ is weakly compatible and θ is the Y-rule.

(b) (P) is unspecified and θ is the M-rule.

(c) (P) is the property that Σ is weakly compatible and θ is the M/Y-combination closure rule $\theta_{M/Y}$ which applies θ_M or θ_Y to Σ.

To elucidate the notion of a split closure sequence and a split closure associated to a set in $\mathcal{P}(X)$ we next present an example for the assignments of (P) and θ specified in (a). Consider the set $X = \{1, 2, 3, 4, 5\}$ together with the collection Σ comprising of the partial X-splits $S_1 = 12|34$, $S_2 = 23|14$, $S_3 = 15|24$, and $S_4 = 45|13$. Clearly, Σ is displayed by an X-cycle C with vertex ordering $1, 2, 3, 4, 5$. Thus Σ is circular and so, by Lemma 4.2, Σ is weakly compatible. Now θ_Y applied to $\{S_1, S_2, S_3\}$ generates the split $S_3' = 15|234$, θ_Y applied to $\{S_1, S_2, S_4\}$ generates the split $S_4' = 45|123$ and θ_Y applied to $\{S_2, S_3', S_4'\}$ generates the split $S_2' = 145|23$. Since every subset of $\Sigma' = \{S_1, S_2', S_3', S_4'\}$ of size three contains two pairwise compatible full splits, θ_Y can only be applied trivially to Σ'. Hence, the sequence $S_0 = \Sigma$, $\Sigma_1 = \{S_1, S_2, S_3', S_4\}$, $\Sigma_2 = \{S_1, S_2, S_3', S_4'\}$, Σ' is a split closure sequence for Σ of length 3 and Σ' is a split closure for Σ.

Regarding (c), it should be noted that even if for some $\Sigma \in \mathcal{P}(X)$ two distinct split closure sequences have the same length and terminate in the same element $\Sigma' \neq \omega$ one of them might utilise fewer applications of θ_Y (and thus more applications of θ_M!) than the other. For the previous example, one way to construct two such sequences is to exploit the following relationship between the Y-rule and the M-rule for $\{S_2, S_3', S_4'\}$.

Proposition 5.1 *Suppose* $\Sigma = \{S_i = A_i|\widetilde{A_i} : i = 1, 2, 3\} \in \mathcal{P}(X)$ *is such that* $A_1 \subseteq A_2$ *and* $\widetilde{A_2} - \widetilde{A_1} \subseteq \widetilde{A_3} \subseteq \widetilde{A_1} \cup \widetilde{A_2}$. *If the* Y-rule applies to Σ then

$$\theta_Y(\Sigma)^- = \{\widetilde{A_1} \cup \widetilde{A_2}|A_1, S_2, S_3\} = \{S_3\} \cup \theta_M(S_1, S_2)^-.$$

Proof: Assume that Σ and S_i and A_i, $i = 1, 2, 3$, are such that the assumptions of the proposition are satisfied. Then $A_1 \cap \widetilde{A_2} = \emptyset$. Combined with the assumption that θ_Y applies to Σ, it follows that either (2) is satisfied with A_i, $i = 1, 2, 3$ playing the roles of their namesakes in the statement of (2) or (2) is satisfied with A_3 playing the role of $\widetilde{A_3}$ and A_i playing the role of A_i,

$i = 1, 2$, in that statement. But the latter alternative cannot hold since this implies $A_1 \cap A_2 \cap \widetilde{A_3} \neq \emptyset$ whereas the assumption $\widetilde{A_3} \subseteq \widetilde{A_1} \cup \widetilde{A_2}$ implies

$$A_1 \cap A_2 \cap \widetilde{A_3} \subseteq A_1 \cap A_2 \cap (\widetilde{A_1} \cup \widetilde{A_2}) = (A_1 \cap A_2 \cap \widetilde{A_1}) \cup (A_1 \cap A_2 \cap \widetilde{A_2}) = \emptyset.$$

Hence, (2) is satisfied with A_i, $i = 1, 2, 3$ playing the roles of their namesakes in the statement of (2). Let S_i', $i = 1, 2, 3$ be as in the statement of the Y-rule. Then $\widetilde{A_2} - \widetilde{A_1} \subseteq \widetilde{A_3}$ implies $S_1' = \widetilde{A_1} \cup \widetilde{A_2} | A_1$. And since $A_1 \subseteq A_2$, we have $S_2' = S_2$ and $A_1 \cap \widetilde{A_2} = \emptyset$ which in turn implies $S_3' = S_3$. Consequently,

$$\theta_Y(\Sigma)^- = \{\widetilde{A_1} \cup \widetilde{A_2} | A_1, S_2, S_3\}.$$

To observe the remaining set equality, note that since (2) is satisfied with A_i, $i = 1, 2, 3$ playing the roles of their namesakes in the statement of (2) neither $A_1 \cap A_2$ nor $\widetilde{A_1} \cap \widetilde{A_2}$ can be empty. Let S_1' be as in the statement of the M-rule. Then $A_1 \subseteq A_2$ implies $S_1' = \widetilde{A_1} \cup \widetilde{A_2} | A_1$ and $S_2' = \widetilde{A_1} \cap \widetilde{A_2} | A_2$. This implies the sought after set equality and thus proves the proposition. ∎

Clearly independent of which one of the rules θ_Y, θ_M or $\theta_{M/Y}$ is applied, a split closure sequence must always be finite since X is finite. In addition and by applying the same arguments as Semple and Steel in [21] one can show that, for the assignments of (P) and θ as described in (a), the length of a split closure sequence for a weakly compatible set $\Sigma \in \mathcal{P}(X)$ is bounded from above by $|\Sigma| \cdot |X| - \Sigma_{\{A,B\} \in \Sigma} |A \cup B|$.

5.2 Split closures

We start with a lemma that is crucial for showing that the split closure of some collection $\Sigma \in \mathcal{P}(X)$ is unique in any of the three combinations for (P) and θ stated in (a) – (c).

Lemma 5.2 *Suppose $\Sigma \in \mathcal{P}(X)$, $\overline{\Sigma} \neq \omega$ is a split closure of Σ and Σ_r and Σ_{r+1} are two consecutive elements in a split closure sequence for Σ.*

(i) If Σ_r is weakly compatible, $\Sigma_r \preceq \overline{\Sigma}$ and Σ_{r+1} is obtained from Σ_r by one application of θ_Y, then Σ_{r+1} is weakly compatible and $\Sigma_{r+1} \preceq \overline{\Sigma}$.

(ii) If Σ_{r+1} is obtained from Σ_r by one application of θ_M and $\Sigma_r \preceq \overline{\Sigma}$, then $\Sigma_{r+1} \preceq \overline{\Sigma}$.

Proof: Suppose $\Sigma, \overline{\Sigma}, \Sigma_r, \Sigma_{r+1}$ are as in the statement of the lemma.

(i) Assume Σ_{r+1} is obtained from Σ_r by applying θ_Y to some set $\{S_1, S_2, S_3\}$ contained in Σ_r. For $i = 1, 2, 3$ and with $A_i \in S_i$ playing the role of their namesakes in the statement of (2), we obtain

$$S_3' = A_3 \cup (A_1 \cap \widetilde{A_2}) | \widetilde{A_3}, S_2' = A_2' \cup (A_1 \cap \widetilde{A_3}) | \widetilde{A_2}, \text{ and } S_1' = \widetilde{A_1} \cup (\widetilde{A_3} \cap \widetilde{A_2}) | A_1.$$

It follows that

$$\Sigma_{r+1} = (\Sigma_r \cup \{S_1', S_2', S_3'\})^-.$$

Since $\Sigma_r \preceq \overline{\Sigma}$, there exist partial splits $S_i'' = A_i'' | \widetilde{A_i''} \in \overline{\Sigma}$ with S_i'' extending S_i, $i = 1, 2, 3$. Without loss of generality we may assume for all i that $A_i \subseteq A_i''$ and $\widetilde{A_i} \subseteq \widetilde{A_i''}$. Since $\overline{\Sigma}$ is weakly compatible, (2) is satisfied by $\{S_1'', S_2'', S_3''\}$ with $A_1'' \cap \widetilde{A_2''} \cap \widetilde{A_3''} = \emptyset$. Since, by assumption, $\overline{\Sigma} \neq \omega$ and so θ_Y applies trivially to $\overline{\Sigma}$, we must have $A_1'' \cap \widetilde{A_2''} \subseteq A_3''$, $A_1'' \cap \widetilde{A_3''} \subseteq A_2''$, and $\widetilde{A_3''} \cap \widetilde{A_2''} \subseteq A_1''$. It follows that, for all $i = 1, 2, 3$, S_i' is extended by S_i'' which in turn implies $\Sigma_{r+1} \preceq \overline{\Sigma}$. Since $\overline{\Sigma} \neq \omega$ and so $\overline{\Sigma}$ is weakly compatible, Lemma 4.1 implies that Σ_{r+1} is weakly compatible.

(ii): Suppose Σ_{r+1} is obtained from Σ_r by applying θ_M to some set $\{S_1, S_2\} \subseteq \Sigma_r$. Put $S_i = A_i | \widetilde{A_i}$, $i = 1, 2$, and assume without loss of generality, that $A_1 \cap A_2 \neq \emptyset$ and $\widetilde{A_1} \cap \widetilde{A_2} \neq \emptyset$. Then

$$S_1' = A_1 \cap A_2 | \widetilde{A_1} \cup \widetilde{A_2''} \text{ and } S_2' = A_1 \cup A_2 | \widetilde{A_1} \cap \widetilde{A_2''}$$

and so

$$\Sigma_{r+1} = (\Sigma_r \cup \{S_1', S_2'\})^-.$$

By assumption, $\Sigma_r \preceq \overline{\Sigma}$ and so there exist partial splits $S_1'' = A_1'' | \widetilde{A_1''}$, $S_2'' = A_2'' | \widetilde{A_2''}$ in $\overline{\Sigma}$ with S_i'' extending S_i, $i = 1, 2$. Without loss of generality we may assume $A_i \subseteq A_i''$ and $\widetilde{A_i} \subseteq \widetilde{A_i''}$, $i = 1, 2$. Then $A_1'' \cap A_2'' \neq \emptyset$ and $\widetilde{A_1''} \cap \widetilde{A_2''} \neq \emptyset$. Since, by assumption, $\overline{\Sigma} \neq \omega$ and so θ_M only applies trivially to $\overline{\Sigma}$ we have $\theta_M(S_1'', S_2'') \preceq \overline{\Sigma}$. Hence, there exist partial splits $S_j = A_j | \widetilde{A_j} \in \overline{\Sigma}$, $j = 3, 4$, so that $A_1'' \cap A_2'' | \widetilde{A_1''} \cup \widetilde{A_2''}$ is extended by S_3 and $\widetilde{A_1''} \cap \widetilde{A_2''} | A_1'' \cup A_2''$ is extended by S_4. Without loss of generality, we may assume that $A_1'' \cap A_2'' \subseteq A_3$ and $\widetilde{A_1''} \cup \widetilde{A_2''} \subseteq \widetilde{A_3}$ and that $A_1'' \cup A_2'' \subseteq A_4$ and $\widetilde{A_1''} \cap \widetilde{A_2''} \subseteq \widetilde{A_4}$. Then

$$A_1 \cap A_2 \subseteq A_1'' \cap A_2'' \subseteq A_3 \text{ and } \widetilde{A_1} \cup \widetilde{A_2} \subseteq \widetilde{A_1''} \cup \widetilde{A_2''} \subseteq \widetilde{A_3}$$

and so S_3 extends S_1'. Similarly, it follows that S_4 extends S_2'. Thus, $\Sigma_{r+1} \preceq \overline{\Sigma}$. ∎

With this result in hand, we are now in the position to present a key result.

Theorem 5.3 *Suppose $\Sigma \in \mathcal{P}(X)$. Then any two split closures for Σ are the same if*

(i) Σ is weakly compatible and solely the Y-rule is used to obtain a split closure for Σ,

(ii) solely the M-rule is used to obtain a split closure for Σ, or

(iii) Σ is weakly compatible and solely the M/Y-rule is used to obtain a split closure for Σ.

Proof: Suppose $\Sigma \in \mathcal{P}(X)$. We start with remarking that we prove Statements (i), (ii) and (iii) collectively as the proof of all three statements relies on an inductive argument on the length of a split closure sequence for Σ. However, since the arguments for the inductive step differ under the assumptions made in (i), (ii) and (iii), we discuss each inductive step separately.

Suppose that the assumptions made in (i) or in (ii) or in (iii) hold. If every split closure of Σ is ω then the theorem holds trivially. So we may assume that there exists a split closure $\overline{\Sigma}$ of Σ with $\overline{\Sigma} \neq \omega$. We proceed by showing that every other split closure of Σ must equal $\overline{\Sigma}$. Suppose that $\sigma : \Sigma_0 = \Sigma, \Sigma_1, \Sigma_2, \ldots, \Sigma_n$ is a split closure sequence of Σ. We now use induction on n to show that if Σ satisfies the assumptions made:

in (i) then, for all $i \in \{0, 1, \ldots, n\}$,

$$\Sigma_i \text{ is weakly compatible and } \Sigma_i \preceq \overline{\Sigma}; \tag{5}$$

in (ii) then, for all $i \in \{0, 1, \ldots, n\}$,

$$\Sigma_i \preceq \overline{\Sigma}; \tag{6}$$

in (iii) then, for all $i \in \{0, 1, \ldots, n\}$,

$$\Sigma_i \text{ is weakly compatible and } \Sigma_i \preceq \overline{\Sigma}. \tag{7}$$

We start with assuming that Σ satisfies the assumptions made in (i), that is, Σ is weakly compatible and solely the Y-rule is used to generate the elements of σ. If $i = 0$ then (5) obviously holds since then $\Sigma_i = \Sigma_0$ and Σ_0 satisfies the properties stated in (5). Now suppose that (5) holds for some $i \in \{0, 1, \ldots, n-1\}$. Then, by Lemma 5.2(i), Σ_{i+1} is weakly compatible and $\Sigma_{i+1} \preceq \overline{\Sigma}$. This completes the induction step and thereby establishes (5).

Next, assume that only the M-rule is used to generate the elements in Σ. If $i = 0$, then (6) holds since then $\Sigma_i = \Sigma_0$ and Σ_0 satisfies (6). Assume that (6) holds for some $i \in \{0, 1, \ldots, n-1\}$. Then Lemma 5.2(ii) implies $\Sigma_{r+1} \preceq \overline{\Sigma}$ which completes the induction step and thereby establishes (6).

Finally, assume that Σ satisfies the assumptions made in (iii), that is, Σ_r is weakly compatible and only the M/Y-rule is used to generate the elements in σ. If $i = 0$, then (7) obviously holds since then $\Sigma_i = \Sigma_0$ and Σ_0 is closed under $\theta_{M/Y}$. Now suppose that (7) holds for some $i \in \{0, 1, \ldots, n-1\}$. Then Σ_i is weakly compatible, $\Sigma_r \preceq \overline{\Sigma}$, and one of the following two cases must hold. Either (a) Σ_{r+1} is obtained from Σ_r by applying θ_Y or (b) Σ_{r+1} is obtained from Σ_r by applying θ_M.

If Case (a) holds, the proof of the inductive step in (i) implies that Σ_{i+1} is weakly compatible and $\Sigma_{i+1} \preceq \overline{\Sigma}$.

If Case (b) holds, $\Sigma_{i+1} \preceq \overline{\Sigma}$ follows from the proof of the inductive step in (ii). That Σ_{i+1} is weakly compatible follows from Lemma 5.2(ii) and the fact that $\overline{\Sigma}$ is weakly compatible. This completes the induction step and thereby establishes (7).

We conclude with noting that for $i = n$, we obtain $\Sigma_n \preceq \overline{\Sigma}$ regardless of whether we are assuming (i) or (ii) or (iii) to hold. In case of (i) holding and applying (5) to $i = n$ or (iii) holding and applying (7) to $i = n$, we see that Σ_n is weakly compatible. By interchanging the roles of Σ_n and $\overline{\Sigma}$, we deduce $\overline{\Sigma} \preceq \Sigma_n$. Thus, under the assumptions made in (i) or (ii) or (iii), we have $\overline{\Sigma} = \Sigma_n$ which concludes the proof of the theorem. ∎

Extending in the case of (P) denoting the condition "Σ is weakly compatible" and θ denoting either the Y-rule or the M/Y-rule, the definition of the split closure to non weakly compatible sets in $\mathcal{P}(X)$ by defining the split closure of such sets to be ω, we obtain

Corollary 5.4 *Suppose θ is either the Y- or M- or M/Y-rule. Then for any $\Sigma \in \mathcal{P}(X)$, any two split closures for Σ obtained via θ are the same.*

Bearing in mind Corollary 5.4, we denote for $\theta \in \{\theta_Y, \theta_M, \theta_{M/Y}\}$ the split closure of a set $\Sigma \in \mathcal{P}(X)$ by $\langle \Sigma \rangle_\theta$, Note that $|\langle \Sigma \rangle_{\theta_Y}| \leq |\Sigma|$ but that neither $|\langle \Sigma \rangle_{\theta_M}| \leq |\Sigma|$ nor $|\langle \Sigma \rangle_{\theta_{M/Y}}| \leq |\Sigma|$ have to hold. Also note that, if we denote the collection $\mathcal{P}(X) \cup \{\omega\}$ by $\mathcal{P}_\omega(X)$, define $\langle \omega \rangle_\theta = \omega$ for some closure rule θ, and put $\Sigma \preceq \omega$ for all $\Sigma \in \mathcal{P}(X)$, then the split closure with respect to $\theta \in \{\theta_Y, \theta_M, M/Y\}$ satisfies the usual properties of a closure operation. More precisely, for all $\Sigma, \Sigma' \in \mathcal{P}_\omega(X)$ we have $\Sigma \preceq \langle \Sigma \rangle_\theta$, if $\Sigma \preceq \Sigma'$ then $\langle \Sigma \rangle_\theta \preceq \langle \Sigma' \rangle_\theta$, and $\langle \langle \Sigma \rangle_\theta \rangle_\theta = \langle \Sigma \rangle_\theta$.

As an immediate consequence of Lemma 4.2, Proposition 4.4, and Theorem 5.3, we obtain our main result which we state next.

Corollary 5.5 *Suppose $\Sigma \in \mathcal{P}(X)$ and C is an X-cycle. Then Σ is displayed by C if and only if $\langle \Sigma \rangle_{\theta_{M/Y}}$ is displayed by C. In that case $\langle \Sigma \rangle_{\theta_Y}$ and $\langle \Sigma \rangle_{\theta_M}$ are also displayed by C.*

We conclude this section with remarking that, in general, not all elements in $\langle \Sigma \rangle_{\theta_{M/Y}}$ need to be full splits on X, $\Sigma \in \mathcal{P}(X)$ circular. However, it is reasonable to assume that those that have been extended to full splits on X contain phylogenetically relevant information and programs such as e. g. SplitsTree4 [12] may be employed to produce a circular phylogenetic network that displays them. For the following we refer to the combination of the M/Y-rule with a phylogenetic network generation package such as SplitsTree4 as the *MY-closure approach*. Although a detailed analysis of this approach is beyond the scope of this paper and will be presented elsewhere, we note that the MY-closure approach cannot be polynomial in the worst case since if the collection of partial splits comprises of all $3\binom{n}{4}$ partial X-splits $A|B$ with $|A| = 2 = |B|$ and $n = |X|$ then $\theta_{M/Y}$ will generate all $2^{(n-1)} - 1$ splits of X

6   An example: The ring of life

One of the most fiercely debated questions amongst biologists is the origin of eukaryotes (essentially cells that have a nucleus and organelles) [16]. The main reason for this is that eukaryotes have eubakteria-like genes as well as archaebacteria-like genes making it very difficult to establish the evolutionary relationships between eukaryotes and prokaryotes (essentially cells that lack nucleus and organelles) which is the collective name for eubakteria and archaebacteria. To help shed light into this question, Rivera et al. [20] analysed 10 bacterial genomes. The 5 most probable phylogenetic trees resulting from their analysis are presented in Fig. 1 of that paper. For the convenience of the reader, we depict them in slightly different form in Fig. 5. Note that the collection of splits displayed by these trees is circular and also that, when ignoring the fact that the leaves are marked with different symbols, the last 2 trees are the same.

Using a technique called *Conditioned Reconstruction* [20], Rivera et al. constructed the phylogenetic network depicted in Fig. 6 with the degree 5 interior vertex plus all its incident edges removed and all resulting degree 2 vertices suppressed. The resulting structure they then interpreted as lending
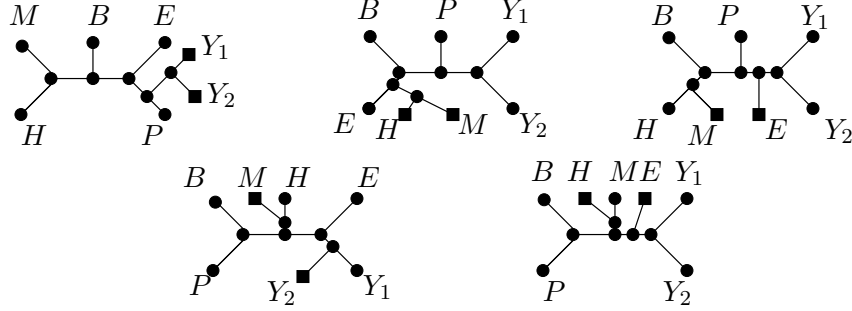
Figure 5: Five most probable phylogenetic trees that appeared in slightly different form in [20, Fig. 1] (cf text for details).

support to the idea that, in its early stages, evolution was not tree-like but rather more like a ring (hence the term "ring of life") with the eukaryotic genome being the result of a fusion of 2 diverse procaryotic genomes [20].

To find out how dependent Rivera et al.'s ring of life is on the fact that all 5 trees are on the same leaf set, we randomly removed pairs of leaves plus their incident edges (suppressing resulting degree 2 vertices and always ensuring that there were no 2 trees from which the same pair of leaves was removed) resulting in 5 trees $T_1, \ldots, T_5$ on 5 leaves. Perhaps not surprisingly, we found that, in general, removal of pairs of leaves did not allow us to recover Rivera et al.'s ring of life. The exception being the trees depicted in Fig. 5 with the leaves marked by a filled-in square removed. For these 5 trees the associated phylogenetic network $\mathcal{N}(T_1, \ldots, T_5)$ produced by the MY-closure approach is depicted in Fig. 6(left).

In addition and with the exception of one instance where one split in $\Sigma(T_1, \ldots, T_5)$, that is, the set of all splits displayed by $T_1, \ldots, T_5$, was not extended to a full split by our closure rules and thus was not displayed by $\mathcal{N}(T_1, \ldots, T_5)$ our rules always generated a minimum collection of splits so that $\mathcal{N}(T_1, \ldots, T_5)$ displayed all the splits in $\Sigma(T_1, \ldots, T_5)$.

Interestingly, both the Z-closure super-network and Q-imputation approach seemed to struggle with this example with, in the case of Z-closure super-network, either yielding a very complex network $\mathcal{N}(T_1, \ldots, T_5)$ in which numerous extensions of one and the same split in $\Sigma(T_1, \ldots, T_5)$ was displayed (see Fig. 6(right)) or $\mathcal{N}(T_1, \ldots, T_5)$ displayed only a subset of splits in $\Sigma(T_1, \ldots, T_5)$ (Q-imputation).
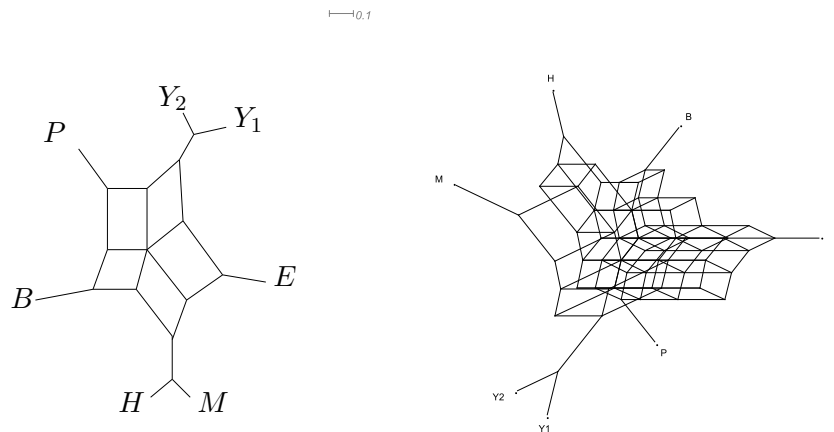
Figure 6: Left, a circular network on 2 yeast genomes, an α-probacterium, a bacillus, a halobacterium, a methnaococcus, an ecocyte, and an archaeoglobium (the genome abbreviations follow [20]). It displays the split collection inferred from the collection of partial splits induced by the trees in Fig. 5 with the leaves marked with a square plus their incident edges removed and the resulting degree 2 vertices suppressed using the MY-closure approach. Right, the Z-closure super-network on the same set of partial splits
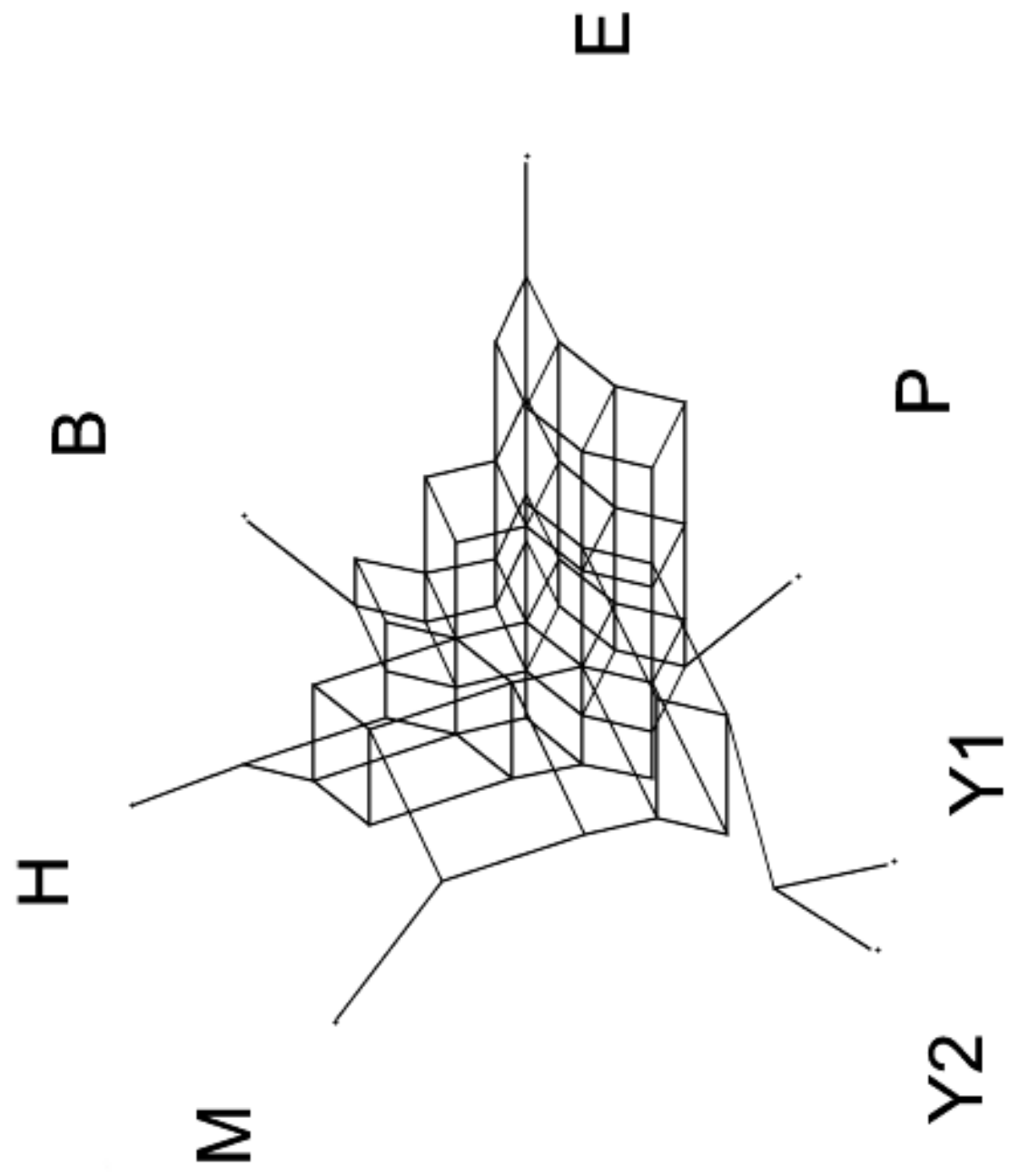
ments.

References

[1] Bandelt,H., Dress A., 1992. A canonical decomposition theory for metrics on a finite set, Adv. Math. 92, 47–105.

[2] Bininda-Emonds, O., 2004. Phylogenetic supertrees: combining information to reveal the tree of life, Computational Biology Series 4, Kluwer.

[3] Bryant, D., Moulton, V., 2004. Neighbor-Net: an agglomerative method for the construction of phylogenetic networks, Mol. Biol. Evol. 21(2), 255-265.

[4] Dress, A., Huson, D. H., 2004. Constructing splits graphs, IEEE/ACM transactions on Computational Biology and Bioinformatics, 1, 109-115.

[5] Fitzpatrick, D., Creevey, C., McInerney, J., 2006. Genome phylogenies indicate a meaningful alpha-proteobacterial phylogeny and support a grouping of the mitochondria with the Rickettsiales, Mol. Biol. Evol., 23, 74 - 85.

[6] Grünewald, S., Forslund, K., Dress, A., Moulton, V.,. QNet: an agglomerative method for the construction of phylogenetic networks from weighted quartets, Mol. Biol. Evol., in press.

[7] Gusfield D., Edduh S., Langley G., The fine structure of galls in phylogenetic networks, to appear in INFORMS J. of Computing Special Issue of Computational Biology.

[8] Hamed, B., M., 2005. Neighbour-nets portray the Chinese dialect continuum and the linguistic legacy of China's demic history, Proc. R. Soc. B, 272(1567), 1471-2954.

[9] Holland, B., Conner, G., Huber, K. T., Moulton, V.,. Imputing supertrees and supernetworks from quartets, Systematic Biology, in press.

[10] Huber, K. T., Moulton, V., 2005. Phylogenetic networks, In: Mathematics of Evolution and Phylogeny, Eds: O. Gascuel, Oxford University Press.

[11] Huber, K. T., Moulton, V., 2006. Phylogenetic networks from multi-labelled trees. J. Math. Biol. 52:613-632.

[12] Huson, D. H., Bryant, D., 2005. Application of phylogenetic networks in evolutionary studies, Mol. Biol. Evol. 23(4), 254-267.

[13] Huson, D. H., Dezulian, T., Kloepper, T. K., Steel, M. A., 2004. Phylogenetic super-networks from partial trees, IEEE Trans. Comp. Biol. Bioinform. 1, 151-158.

[14] Huson, D. H., Kloepper, T. H., Lockhart, P. J., Steel, M. A., 2005. Reconstruction of reticulate networks from gene trees, LNBI 3500, p 233-249

[15] Lake, J. A., Rivera, M. C., 2004. Deriving the genomic tree of life in the presence of horizontal gene transfer: Conditioned Reconstruction, Mol. Biol. Evol. 21 681-690.

[16] Martin, W., Embley, T. M., 2004. Early evolution comes full circle, Nature, 431, 134-137.

[17] McBreen, K., Lockhart, P. J.,. Reconstructing reticulate evolutionary histories in plants, submitted.

[18] Meacham, C.A., The recovery of trees from measures of dissimilarity, In: Mathematics in the Archaeological and Historical Sciences, F.R. Hodson, D. G. Kendall, and P. Tatu (eds), Edinburgh University Press (1971), 387- 395.

[19] Morrison, D., 2005. Networks in phylogenetic analysis: new tools for population biology, International Journal of Parasitology, 35 567-582.

[20] Rivera, M. C., Lake, J.A., 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes, Nature, 43, 152-155.

[21] Semple, C., Steel, M., Tree reconstruction via closure operations on partial splits, JOBIM 2000, LNCS 2066 (2001) 126-134.

[22] Semple, C., Steel, M., Phylogenetics, Oxford University Press, 2003.

[23] Simonson, A. B., Servin, J. A., Skophammer, R. G., Herbold, C. W., Rivera, M. C., Lake, J. A., 2005. Decoding the genomic tree of life, PNAS, 102(suppl 1), 6608-6613.

B

E

H

M

P

Y1

Y2

(b)